# PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-148770
(43)Date of publication of application : 30.05.2000

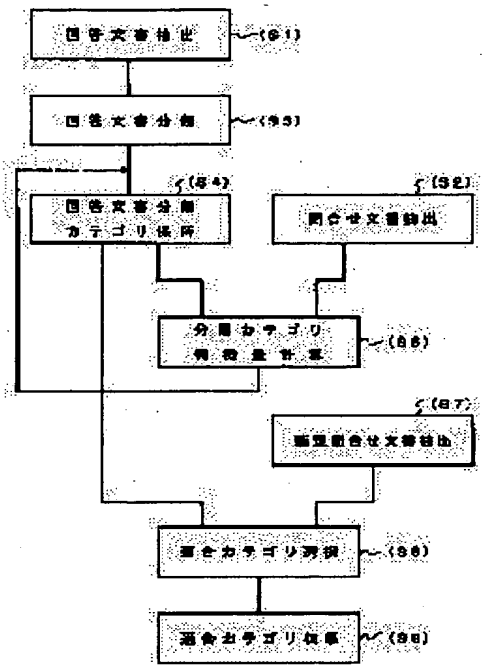
(51)Int.Cl. G06F 17/30

(21)Application number : 10-315625
(22)Date of filing : 06.11.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : MORI DAIJIRO
OKUBO MASAKATSU
SUGIZAKI MASAYUKI
TANAKA KAZUO

## (54) DEVICE AND METHOD FOR CLASSIFYING QUESTION DOCUMENTS AND RECORD MEDIUM WHERE PROGRAM WHEREIN SAME METHOD IS DESCRIBED IS RECORDED

(57)Abstract:
PROBLEM TO BE SOLVED: To retrieve a category matching the question contents of a question document with high precision by using a feature quantity extracted from constituent elements of the question document corresponding to an answer document included in the category when the feature quantity of the extracted category is calculated.
SOLUTION: Individual documents are extracted from an answer document set and documents are put together form a question document set (S1, S2). Those individual documents are classified (S3) after they are decomposed into words and morpheme analysis is carried out to calculate feature vectors. After the classified answer documents are held (S4), feature vectors are calculated as to the question documents and classification category feature quantities are calculated (S6) while made to correspond to the answer documents. Then a new question document is extracted (S7), a feature vector is calculated as to the new document to calculate the adaptivity to the held classification category (S8), and a specific number of matching categories from the top are gathered (S9).



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

---

CLAIMS

---

[Claim(s)]
[Claim 1] Equipment which matches and manages two or more inquiry documents characterized by providing the following, and the reply document to each inquiry A reply document classification means to classify a reply document set into two or more categories according to the characteristic quantity from which it is extracted from each reply document A classification category characteristic quantity calculation means to calculate the characteristic quantity of this category using the characteristic quantity of the inquiry document corresponding to each reply document contained in each category classified according to the aforementioned reply document classification means, A conformity category selection means by which the characteristic quantity obtained by the aforementioned classification category characteristic quantity calculation means out of the category which was given newly, and which asked and was classified according to the aforementioned reply document classification means to the document extracts what is suited with the characteristic quantity of this new inquiry document
[Claim 2] How to match and manage two or more inquiry documents characterized by providing the following, and the reply document to each inquiry The reply document hierarchy which classifies a reply document set into two or more categories according to the characteristic quantity from which it is extracted from each reply document The classification category characteristic quantity calculation stage which calculates the characteristic quantity of this category using the characteristic quantity of the inquiry document corresponding to each reply document contained in each category classified according to the aforementioned reply document hierarchy, The conformity category selection stage where the characteristic quantity obtained according to the aforementioned classification category characteristic quantity calculation stage out of the category which was given newly, and which asked and was classified according to the aforementioned reply document hierarchy to the document extracts what is suited with the characteristic quantity of this new inquiry document
[Claim 3] In the record medium which described how to match and manage two or more inquiry documents and the reply document to each inquiry, in the form of a program, and recorded the program concerned The reply document hierarchy according to which the aforementioned method of carrying out management classifies a reply document set into two or more categories according to the characteristic quantity from which it is extracted from each reply document, The classification category characteristic quantity calculation stage which calculates the characteristic quantity of this category using the characteristic quantity of the inquiry document corresponding to each reply document contained in each category classified according to the aforementioned reply document hierarchy, Out of the category which was given newly and which asked and was classified according to the aforementioned reply document hierarchy to the document The record medium characterized by having equipped the characteristic quantity obtained according to the aforementioned classification category characteristic quantity calculation stage with the characteristic quantity of this new inquiry document, and the conformity category selection stage of extracting what suiting, having described the method concerned in the form of a program, and recording the program concerned.

---

## DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] In the business which answers to a lot of inquiries, this invention extracts the combination to which the content is similar out of the past inquiry and reply history, and relates to the record medium which recorded the program which described the method concerned on the classification equipment and the method row of an inquiry document which perform support or automation of reply work.
[0002]
[Description of the Prior Art] the document with which the content is similar is extracted out of a lot of accumulation document set with development of natural-language-processing technology, and improvement in the throughput of a computer, and it classifies into two or more categories -- things are possible The technology of selecting a category with the highest goodness of fit from the existing categories about the document given newly is also well-known.
[0003] The following technique is known as the classification method of a document set. First, the document used as the candidate for a classification is disassembled into the element which makes a character string, a word, and a clause a unit, and characteristic quantity is calculated based on the combination of this element. Next, it asks for the degree of similar of characteristic quantity, and the combination of all documents constitutes a cluster in an order from combination with the high degree of similar. Let this cluster be a ****** classification result repeatedly until it reaches a fixed goodness of fit or the size of a cluster in this process.
[0004] Various methods as the calculation method of characteristic quantity are devised. For example, after disassembling a document into the element which makes a character string, a word, and a clause a unit as mentioned above, it asks for the weight of an element based on the frequency of occurrence in a document set of each element, and the frequency of occurrence in this document, and the method of expressing characteristic quantity by the vector constituted by each element and its weight is learned.
[0005] Or an element is suitably arranged on n-dimensional vector space, and how to calculate the feature vector of a document by the sum of the vector which the element contained in a document makes is also learned so that the degree of association between the elements contained in a document may be computed by the predetermined method and the high element of degree of association may serve as near.
[0006] As a method of calculating the degree of similar between characteristic quantity, when characteristic quantity is expressed as a vector, the method of computing a goodness of fit by the inner product or cosine which two vectors accomplish is used widely.
[0007]
[Problem(s) to be Solved by the Invention] The document itself which is all a candidate for a classification is made into the information source, and it is classifying according to a Prior art based on the element which constitutes this document. However, in the inquiry document which receives from many and unspecified men, since the vocabulary and expression which are used change with people, even if the content of an inquiry is the same, the components of a document may differ.
[0008] For this reason, it is difficult to perform the classification adapted to the content of an inquiry with high precision to an inquiry document set.
[0009] this invention was made in view of the technical problem looked at by Prior art which was mentioned above, and aims at performing the classification based according to the content of an inquiry.
[0010]
[Means for Solving the Problem] the document itself which is all a candidate for a classification is made into the information source, and it classifies according to this invention to having classified based on the element which constitutes this document based on the element which constitutes the reply document corresponding to this inquiry document, and in calculating the characteristic quantity of each category obtained as a classification result, an inquiry document consists of Prior arts -- the element is used

[0011] As mentioned above, in an inquiry document, since the vocabulary and expression which are used are various, the characteristic quantity extracted from there may not agree with the contents of an inquiry.
[0012] In the business answered to the inquiry received from many and unspecified men on the other hand Since the number of people which draws up a reply document is smaller and it is unified in many cases among respondents about terminological use compared with the number of those who draw up an inquiry document, The vocabulary which appears in a reply document, and expression are more uniform compared with the vocabulary and expression which appear in an inquiry document, and its inclination for the reply document using the same vocabulary and same expression to be drawn up to the same contents of an inquiry is strong. Therefore, though the vocabulary contained in an inquiry document is various, if the contents of an inquiry are the same, it is expectable that the characteristic quantity extracted from the reply document corresponding to it shows the high degree of similar.
[0013] Moreover, since the vocabulary which asks a reply document and is used in written form differs from an expressional inclination, Although the category which suits a new inquiry document cannot be searched using the characteristic quantity extracted from the reply document, in this invention Since the characteristic quantity extracted from the component of the inquiry document corresponding to the reply document contained in this category is used in case the characteristic quantity of the extracted category is calculated, a new inquiry document can be considered as an input and KAKODERI which suits the contents of an inquiry of this inquiry document can be searched with high precision.
[0014]
[Embodiments of the Invention] The example of realization of the classification equipment of the inquiry document of this invention is explained.
[0015] The system configuration of this example of realization is shown in drawing 1 .
[0016] The sign 1 in drawing is a reply document set, and the meeting of the past reply document and 2 are inquiry document sets. The meeting of the past inquiry document, What 3 is a reply document classification means, and classifies the contents of the reply document set 1 after calculating a feature vector by having performed morphological analysis about the contents of the reply document set 1, 4 -- in a classification category characteristic quantity calculation means and 7, a new inquiry document and 8 express the conformity category selection means, and, as for a reply document classification category and 5, 9 expresses [ classification category characteristic quantity and 6 ] the conformity category selection result Hereafter, each component shown in drawing 1 is explained in order.
[0017] The flow shown in drawing 2 realizes the reply document classification means 3.
[0018] In drawing 2 , signs 1, 3, and 4 correspond to drawing 1 , in the morphological analysis section and 32, the reply document feature-vector calculation section and 33 express the list for a classification, and 34 expresses [ 31 ] the similar document extraction section.
[0019] As shown in drawing 2 , about each content of the reply document set 1, morphological analysis is performed in the morphological analysis section 31, each sentence of a reply document is decomposed into a word, and a feature vector FV is calculated based on each word in the reply document feature-vector calculation section 32. The result is held as a list 33 for a classification. Subsequently, in the similar document extraction section 34, as the degree of similar is calculated by taking out two elements under list 33 for a classification, combination with the large degree of similar is extracted and a cluster is obtained, the reply document classification category 4 is obtained. Furthermore, it states concretely.
[0020] The morphological analysis section 31 decomposes each sentence in a reply document into a word using well-known technology. The reply document feature-vector calculation section 32 calculates a feature vector FV based on each word contained in a reply document.
[0021]
FV (i) =(w (i, 1), ..., w (i, j), ..., w (i, N)) w(i, j) =tf(i, j) * log here (M/df (j)) i A publication number and N The total number of words in a reply document set, and tf (i, j) Document i Word "j" which can be set The number of times of an appearance, and M The total of a reply document, and df (j) Word in a document set "j" It is the number of times of an appearance.
[0022] Document i Feature vector FV (i) It is expressed as a vector on space with a number of each word of dimensions which constitute a document set. The element of each dimension shows which [ element of the dimension and which / strong ] the document has a relation with. tf (i, j) A term is Document i. Word "j" which can be set Although it is the number of times of an appearance, this means positioning that it is the important word which shows the feature of the document, if the word comes out repeatedly in a document.
[0023] The term of log (M/df (j)) is the word "j". Although it is the number which took the opposite numeric value of the inverse number of the frequency of occurrence, it means interpreting this as it being such an important word that the frequency in which the word appears through the whole document being low. However, if it considers as it is that the inverse number of the frequency of occurrence is significance, it

will be said that the word whose frequency of occurrence is 1/100 is 100 times more important, and since it is unnatural, the logarithm has been taken. In addition, this calculation method is tf* idf which G.Salton developed. It is based on a method and, generally is widely used in text reference.

[0024] In addition, although the vector which uses a word at large as an element is expressing characteristic quantity in this example, how to use only an independent word as an element as how to choose the component of a vector including the method, affix, and compound which are used as an element, and the method of using as an element the word train included in a noun phrase are also considered. In addition, although the calculation method of various characteristic quantity thinks, if it is possible to calculate the degree of similar between these quantity, it is possible [ it is the quantity computed based on the component of a document and ] to apply this invention also in which method.

[0025] The similar document extraction section 34 picks out two elements from the list 33 for a classification, and calculates the degree of similar about all combination.

[0026] Element i Element k The degree of similar R (i, k) It asks by the following formulas.

[0027]

R(i, k) = FV(i) and FV(k)/FV (| FV(i) |·| FV (k) |) (i) FV (k) is a vector which shows the feature of a document, respectively. FV(i) and FV (k) mean the inner product of a vector, and | FV(i) |·| FV(k) | means the product of each magnitude of a vector. That is, R (i, k) expresses the cosine of two vectors. Therefore, it is in the idea that the degree of similar is high, so that the angle which two vectors constitute is small.

[0028] In an initial state, all reply documents are considered as the list for a classification. Cluster C which extracts combination with the largest degree of similar, and is constituted from an element of this combination in the list for a classification It generates. cluster C an element -- i and k it was -- the time -- one cluster C A feature vector is calculated by the following formulas.

[0029] FV (C) = (u(i) * FV (i)+u(k) * FV (k))/(u (i) +u (k)) here, u (i) is i. They are 1 and i if it is a document. It is i if it is a cluster. It considers as the number of the documents contained.

[0030] This is searching for the center of gravity of a feature vector. Since a cluster is the settlement containing two or more documents, only several document minutes contained there, it attaches weight and is searching for the center of gravity of a vector.

[0031] Next, the list 33 for a classification to i k It removes and is Cluster C. It adds, and again, combination with the largest degree of similar is extracted in the list for a classification, and a cluster is generated.

[0032] In this way, the above-mentioned process is repeated until the maximum of the degree of similar is less than a predetermined threshold. When a repeat is completed, the clusters used as the element of other clusters are enumerated, and a set of the document contained in each cluster is considered as a similar document set.

[0033] The flow shown in drawing 3 realizes the classification category characteristic quantity calculation means 6.

[0034] In drawing 3 , signs 2, 4, and 6 correspond to drawing 1 , in an inquiry document and the reply document correspondence Management Department, and 62, the classification category feature-vector calculation section and 63 express the classification category feature vector, and 64 expresses [ 61 ] the morphological analysis section.

[0035] An inquiry document and the reply document correspondence Management Department 61 acquire the inquiry document corresponding to this reply document from the reply document classification category 4 about all the reply documents contained in each classification category. It manages in the state where asked each sentence of this inquiry document and it decomposed into the word through the document set 2 and the morphological analysis section 64. The classification category feature-vector calculation section 62 asks for the following feature vectors FV, and makes this the ****** classification category feature vector 63.

[0036]

[Equation 1]

$$FV(i) = (w(i,1),\ldots,w(i,j),\ldots,w(i,N))$$

$$w(i,j) = \sum_{c=1}^{C(i)} tf(c,j) \quad \log (M / df(j))$$

[0037] Here, it is i. The number of a classification category, and N The total number of words in an inquiry document set and C (i) Classification category i The number of reply documents and tf (c, j) which are contained Reply document c contained in a classification category Word in a corresponding inquiry document "j" The number of times of an appearance, and M The total of an inquiry document, and df (j)

Word in an inquiry document set "j" It is the number of times of an appearance.
[0038] The classification category is called for as a meeting of the document with which plurality is similar. Since the feature vector of this classification category is expressed, the information on the word which the document contained in it contains is used. Although it is the same view as asking for the feature vector of an individual document fundamentally, since two or more documents are included, it is asking for the feature vector of the whole category by adding all the number of times of an appearance of the word contained in each document.
[0039] The flow shown in drawing 4 realizes the conformity category selection means 8.
[0040] drawing 4 -- setting -- signs 4, 7, and 8 -- drawing 1 -- corresponding -- 81 -- in a new inquiry document feature vector and 84, the goodness of fit calculation section and 85 express the conformity category list, and 86 expresses [ the morphological analysis section and 82 / the new inquiry document characteristic quantity calculation section and 83 ] the sorting application section
[0041] The morphological analysis section 81 takes out each sentence of the inquiry document given newly from the new inquiry document 7, and decomposes it into a word.
[0042] In the new inquiry document characteristic quantity calculation section 82, the following inquiry document feature vectors QV are computed as a new inquiry document feature vector 83 based on the word information extracted by the aforementioned morphological analysis section 81.
[0043] QV = (tf (1), ..., tf (j), ..., tf (N)) here, it is tf (j). Word in a new inquiry document "j" The number of times of an appearance, and N It is the total number of words in an inquiry document set.
[0044] The goodness of fit calculation section 84 extracts the feature vector obtained by the aforementioned new inquiry document characteristic quantity calculation section 82 and all the classification categories obtained from the aforementioned classification category characteristic quantity calculation means 6 from the reply document classification category 4, and is the goodness of fit score with the feature vector of the classification category concerned. It calculates by the following formulas and the conformity category list 85 is obtained.
[0045]
score(i) =QV·FV (i) / Here (|QV| and |FV(i) |), QV is the feature vector of a new inquiry document, and FV (i). Classification category i It is a feature vector.
[0046] It is going to ask for the classification category to which many words have lapped with the new inquiry document. Therefore, it is score once decomposing QV into the vector which consists of a word. It has calculated. score It has calculated as a cosine of the feature vector of each classification category, and the feature vector of a new inquiry document to accomplish.
[0047] The goodness of fit calculated by the aforementioned goodness of fit calculation section 84 rearranges in an order from a high thing, extracts a predetermined number of classification categories from a high order, and outputs the sorting application section 86 as a reference result.
[0048] In the above, although the classification equipment of an inquiry document was explained, the program which could take into consideration the classification method of the inquiry document corresponding to the classification equipment of the inquiry document concerned, and described the classification method of the inquiry document concerned again is prepared, and the program concerned can be recorded on a record medium.
[0049] Drawing 5 shows the processing flow showing the classification method of the inquiry document by this invention. The step (S1) or the step (S4) and step (S6), or step (S9) in drawing 5 corresponds to processing of each means shown in the sign 1 or the sign 4 and the sign 6, or sign 9 shown in above-mentioned drawing 1 . Namely, a step (S1): Extract each document from the reply document set 1.
[0050] Step (S2): Extract a document from the inquiry document set 2.
[0051] Step (S3): Classify after calculating a feature vector by having decomposed into the word about each document and having performed morphological analysis.
[0052] Step (S4): Hold the classified reply document.
[0053] Step (S6): Calculate a feature vector about an inquiry document, and calculate classification category characteristic quantity by matching with a reply document. And it returns and holds to a step (S4).
[0054] Step (S7): Extract a new inquiry document.
[0055] Step (S8): Calculate a feature vector about a new inquiry document, and calculate a goodness of fit with the classification category currently held in a step (S6).
[0056] Step (S9): Only a predetermined number collects the categories which suited from a high order.
[0057] The inquiry classification method shown in above-mentioned drawing 5 can describe it in the form of a program, and can record the program concerned on a record medium. Therefore, this invention is made to also include the recorded record medium concerned in the technical range.
[0058]

[Effect of the Invention] As explained above, according to this invention, the accumulated inquiry and reply document information are classified into two or more categories based on the content, and the content becomes possible [ selecting the nearest category in a high precision ] to the newly brought-near inquiry.

[0059] For example, at the inquiry reception window of a company, when the example answered before and a similar inquiry are brought near, the example when replying to last time can be extracted with high precision. However, it replies to often very a lot of inquiries, and the case where extensive existence also of the similar example is recognized very much is possible at the inquiry correspondence window of a company. Since it becomes an operator's burden, a once similar example is summarized as a category and it is made to show an operator by making this category into a reference result in having shown the similar example from one end then. The effect which a work burden mitigates by this since an operator can select the right information out of a fewer number of candidates is acquired.

---

[Translation done.]

---

## DESCRIPTION OF DRAWINGS

---

[Brief Description of the Drawings]
[Drawing 1] It is the system block view of the inquiry document classification equipment of this invention.
[Drawing 2] It is the processing flow of the reply document classification means of the inquiry document classification equipment of this invention.
[Drawing 3] It is the processing flow of the classification category characteristic quantity calculation means of the inquiry document classification equipment of this invention.
[Drawing 4] It is the processing flow of the conformity category selection means of the inquiry document classification equipment of this invention.
[Drawing 5] It is the processing flow of the inquiry document classification method of this invention.
[Description of Notations]
1 Reply Document Set
2 Inquiry Document Set
3 Reply Document Classification Means
4 Reply Document Classification Category
5 Classification Category Characteristic Quantity
6 Classification Category Characteristic Quantity Calculation Means
7 New Inquiry Document
8 Conformity Category Selection Means
9 Conformity Category Selection Result

---

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－148770
（P2000－148770A）
(43)公開日　平成12年５月30日(2000.5.30)

| (51)Int.Cl.⁷ | 識別記号 | ＦＩ | | テーマコード（参考） |
|---|---|---|---|---|
| Ｇ０６Ｆ　17/30 | | Ｇ０６Ｆ　15/401 | ３１０Ｄ | ５Ｂ０７５ |
| | | 15/40 | ３７０Ａ | |

審査請求　未請求　請求項の数３　ＯＬ　（全８頁）

(21)出願番号　特願平10－315625

(22)出願日　平成10年11月６日(1998.11.6)

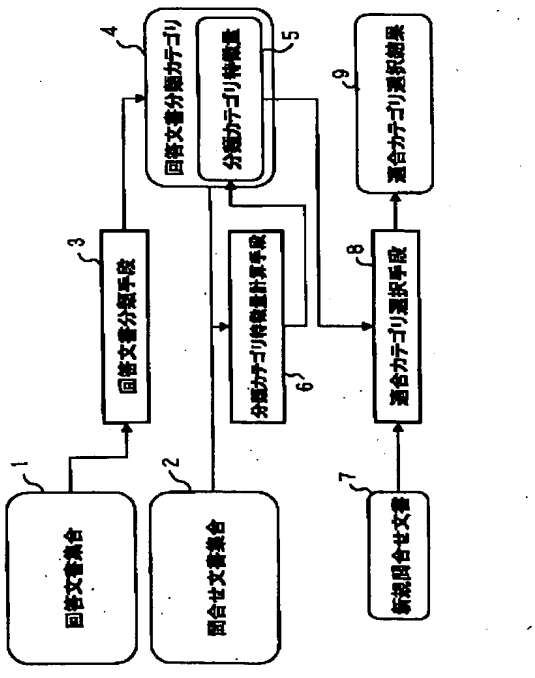
(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　森　大二郎
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(72)発明者　大久保　雅且
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　100087848
弁理士　小笠原　吉義　（外１名）

最終頁に続く

(54)【発明の名称】　問合せ文書の分類装置および方法ならびに当該方法を記述したプログラムを記録した記録媒体

(57)【要約】
【課題】　本発明は、問合せ文を、既存の分類済みの文と対応させて分類してゆくに当って、効率よく高精度で分類することを目的としている。
【解決手段】　問合せ文を、既存の分類済みの文と対応させて分類するに当って、それ以前の問合せ文に対して回答文を作成してその回答文の集合を利用するように構成する。

実施例システムブロック図



1

【特許請求の範囲】

【請求項１】　複数の問合せ文書と、各問合せに対する回答文書とを対応付けて管理する装置において、

回答文書集合を、各回答文書から抽出される特徴量に応じて複数のカテゴリに分類する、回答文書分類手段と、

前記回答文書分類手段によって分類された各カテゴリに含まれる各回答文書に対応する問合せ文書の特徴量を用いて、該カテゴリの特徴量を計算する、分類カテゴリ特徴量計算手段と、

新規に与えられた問合せ文書に対して、

前記回答文書分類手段により分類されたカテゴリの中から、前記分類カテゴリ特徴量計算手段により得られた特徴量が該新規問合せ文書の特徴量と適合するものを抽出する、適合カテゴリ選択手段と、

を備えることを特徴とする問合せ文書の分類装置。

【請求項２】　複数の問合せ文書と、各問合せに対する回答文書とを対応付けて管理する方法において、

回答文書集合を、各回答文書から抽出される特徴量に応じて複数のカテゴリに分類する、回答文書分類段階と、

前記回答文書分類段階によって分類された各カテゴリに含まれる各回答文書に対応する問合せ文書の特徴量を用いて、該カテゴリの特徴量を計算する、分類カテゴリ特徴量計算段階と、

新規に与えられた問合せ文書に対して、

前記回答文書分類段階により分類されたカテゴリの中から、前記分類カテゴリ特徴量計算段階により得られた特徴量が該新規問合せ文書の特徴量と適合するものを抽出する、適合カテゴリ選択段階と、

を備えることを特徴とする問合せ文書の分類方法。

【請求項３】複数の問合せ文書と、各問合せに対する回答文書とを対応付けて管理する方法をプログラムの形で記述して当該プログラムを記録した記録媒体において、

前記管理する方法が回答文書集合を、各回答文書から抽出される特徴量に応じて複数のカテゴリに分類する、回答文書分類段階と、

前記回答文書分類段階によって分類された各カテゴリに含まれる各回答文書に対応する問合せ文書の特徴量を用いて、該カテゴリの特徴量を計算する、分類カテゴリ特徴量計算段階と、

新規に与えられた問合せ文書に対して、前記回答文書分類段階により分類されたカテゴリの中から、前記分類カテゴリ特徴量計算段階により得られた特徴量が該新規問合せ文書の特徴量と適合するものを抽出する、適合カテゴリ選択段階と、

を備え、

当該方法をプログラムの形で記述して、当該プログラムを記録したことを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、大量の問合せに対して回答を行う業務において、過去の問合せ・回答履歴の中から内容が類似する組合せを抽出し、回答作業の支援ないしは、自動化を行う問合せ文書の分類装置および方法ならびに当該方法を記述したプログラムを記録した記録媒体に関する。

2

【０００２】

【従来の技術】自然言語処理技術の発達と計算機の処理能力の向上とに伴い、大量の蓄積文書集合の中から、内容が類似する文書を抽出し、複数のカテゴリに分類することが可能となっている。新規に与えられた文書について、既存のカテゴリの中から最も適合度が高いカテゴリを選び出す技術も公知のものとなっている。

【０００３】文書集合の分類方法としては、以下の手法が知られている。まず、分類対象となる文書を、文字列や単語や文節を単位とする要素に分解し、該要素の組合せに基づいて特徴量を計算する。次に、全ての文書の組合せについて、特徴量の類似度を求め、類似度の高い組合せから順番にクラスタを構成する。この過程を一定の適合度あるいはクラスタのサイズに達するまで繰り返し、該クラスタを以って分類結果とする。

【０００４】特徴量の計算方法としては様々な方式が考案されている。例えば、前述のように文書を、文字列や単語や文節を単位とする要素に分解した後に、各要素の文書集合における出現頻度と該文書における出現頻度とに基づいて要素の重みを求めて、各要素とその重みによって構成されるベクトルによって特徴量を表現する方法が知られている。

【０００５】あるいは、文書に含まれる要素間の関連度を所定の方法で算出し、関連度の高い要素が近傍となるように、ｎ次元のベクトル空間上に要素を適宜配置し、文書に含まれる要素のなすベクトルの和によって文書の特徴ベクトルを計算する方法も知られている。

【０００６】特徴量の間の類似度を計算する方法としては、ベクトルとして特徴量が表現される場合においては、２つのベクトルの成す内積あるいは余弦によって適合度を算出する方法が広く用いられている。

【０００７】

【発明が解決しようとする課題】従来の技術では、いずれも、分類対象である文書そのものを情報源とし、該文書を構成する要素に基づいて分類を行っている。しかし、不特定多数の人から受ける問合せ文書においては、使用される語彙や表現が人によって異なるため、問合せの内容が同一であっても、文書の構成要素が異なる場合がある。

【０００８】このため、問合せ文書集合に対して、問合せの内容に即した分類を高精度に行うことが困難となっている。

【０００９】本発明は、上述したような従来の技術に見られる課題に鑑みてなされたもので、問合せの内容により即した分類を行うことを目的とする。

【００１０】

【課題を解決するための手段】従来の技術では、いずれも、分類対象である文書そのものを情報源とし、該文書を構成する要素に基づいて分類を行っていたのに対して、本発明では、該問合せ文書に対応する回答文書を構成する要素に基づいて分類を行い、分類結果として得られた各カテゴリの特徴量を計算するにあたっては、問合せ文書を構成する要素を用いている。

【００１１】前述のように、問合せ文書においては、使用される語彙や表現が多様であるため、そこから抽出される特徴量が、問合せ内容と合致しない場合がある。

【００１２】一方、不特定多数の人から受ける問合せに回答する業務においては、問合せ文書を作成する人の数に比べて、回答文書を作成する人の数の方が小さく、また用語の使用について回答者の間で統一されている場合が多いため、回答文書に現れる語彙や表現は、問合せ文書に現れる語彙や表現と比べてより一様であり、同一の問合せ内容に対しては、同一の語彙や表現を用いた回答文書が作成される傾向が強い。従って、問合せ文書に含まれる語彙がまちまちであったとしても、その問合せ内容が同一であれば、それに対応する回答文書から抽出される特徴量は高い類似度を示すことが期待できる。

【００１３】また、回答文書と問合せ文書とでは、使用される語彙や表現の傾向が異なるため、回答文書から抽出された特徴量を用いて、新規の問合せ文書に適合するカテゴリを検索することはできないが、本発明では、抽出されたカテゴリの特徴量を計算する際には、該カテゴリに含まれる回答文書に対応する問合せ文書の構成要素から抽出された特徴量を用いるため、新規の問合せ文書を入力とし、該問合せ文書の問合せ内容に適合するカテゴリを高精度に検索することができる。

【００１４】

【発明の実施の形態】本発明の問合せ文書の分類装置の実現例について説明する。

【００１５】本実現例のシステム構成を図１に示す。

【００１６】図中の符号１は回答文書集合であって過去の回答文書の集まり、２は問合せ文書集合であって過去の問合せ文書の集まり、３は回答文書分類手段であって回答文書集合１の内容について形態素解析を行って特徴ベクトルを計算した上で回答文書集合１の内容を分類するもの、４は回答文書分類カテゴリ、５は分類カテゴリ特徴量、６は分類カテゴリ特徴量計算手段、７は新規問合せ文書、８は適合カテゴリ選択手段、９は適合カテゴリ選択結果を表わしている。以下、図１に示される各構成要素について順に説明してゆく。

【００１７】回答文書分類手段３は、図２に示すフローによって実現する。

【００１８】図２において、符号１，３，４は図１に対応し、３１は形態素解析部、３２は回答文書特徴ベクトル計算部、３３は分類対象リスト、３４は類似文書抽出部を表わしている。

【００１９】図２に示す如く、回答文書集合１の個々の内容について、形態素解析部３１にて形態素解析を行って回答文書の各文を単語に分解し、回答文書特徴ベクトル計算部３２にて各単語に基づいて特徴ベクトルFVを計算する。その結果が、分類対象リスト３３として保持される。次いで、類似文書抽出部３４にて、分類対象リスト３３中の２つの要素を取り出して類似度を計算してゆき、類似度の大きい組合せを抽出して、クラスタを得るようにして、回答文書分類カテゴリ４を得る。更に具体的に述べる。

【００２０】形態素解析部３１は、回答文書中の各文を、周知の技術を用いて単語に分解する。回答文書特徴ベクトル計算部３２は、回答文書に含まれる各単語に基づいて、特徴ベクトルFVを計算する。

【００２１】

$FV(i) = (w(i,1), \ldots, w(i,j), \ldots, w(i,N))$

$w(i,j) = tf(i,j) * \log ( M / df(j))$

ここで、i は文書番号、N は回答文書集合における全単語数、tf(i,j) は文書i における単語j の出現回数、M は回答文書の総数、df(j) は文書集合における単語j の出現回数である。

【００２２】文書i の特徴ベクトルFV(i) は、文書集合を構成する各単語の数だけの次元を持つ空間上のベクトルとして表現される。各次元の要素は、その文書がその次元の要素とどれだけ強い関係を持っているのかを示す。tf(i,j) の項は、文書i における単語j の出現回数であるが、これは、その単語が文書中に何度も繰り返し出て来れば、その文書の特徴を示す重要な単語であると位置付けることを意味する。

【００２３】log ( M / df(j))の項は、単語j の出現頻度の逆数の対数値を取った数であるが、これは、その単語が文書全体を通して出現する頻度が少ない程重要な単語だと解釈することを意味する。ただし、出現頻度の逆数をそのまま重要度とみなすと、出現頻度が１００分の１である単語は１００倍重要であるということになり不自然であるため、対数を取っている。なお、この計算方法は、G. Saltonの開発したtf* idf 法によるものであり、テキスト検索においては広く一般的に使用されている。

【００２４】なお、本実施例では、単語全般を要素とするベクトルによって特徴量を表現しているが、ベクトルの構成要素の選び方としては、自立語のみを要素とする方法や、接辞や複合語を含めて要素とする方法、名詞句に含まれる単語列を要素とする方法も考えられる。その他にも様々な特徴量の計算方法が考えるが、文書の構成要素に基づいて算出される数量であり、また、該数量の間の類似度を計算することが可能であれば、いずれの方法においても本発明を適用することが可能である。

【００２５】類似文書抽出部３４は、分類対象リスト３

5

3から2つの要素を取り出して全ての組合せについて類似度を計算する。

【００２６】要素i と要素k との類似度 R(i,k) は、以下の式により求める。

【００２７】

R(i,k)＝ FV(i)・ FV(k)／（｜ FV(i)｜・｜ FV(k)｜）

FV(i)と FV(k)とはそれぞれ文書の特徴を示すベクトルであり、 FV(i)・ FV(k)はベクトルの内積を、｜ FV(i)｜・｜ FV(k)｜はそれぞれのベクトルの大きさの積を意味している。すなわち、R(i,k)は二つのベクトルの余弦を表している。二つのベクトルの成す角が小さいほど類似度が高い、という考えに依っている。

【００２８】初期状態では、全ての回答文書を分類対象リストとする。分類対象リストの中で、最も類似度の大きい組合せを抽出し、該組合せの要素から構成するクラスタC を生成する。クラスタC の要素がi,k であった時、１クラスタC の特徴ベクトルを以下の式により計算する。

【００２９】FV(C) ＝(u(i) * FV(i) ＋u(k)* FV(k))／(u(i) ＋u(k))

ここで、u(i)は、i が文書であれば１、i がクラスタであればi に含まれる文書の数とする。

【００３０】これは、特徴ベクトルの重心を求めている。クラスタは複数の文書を含むまとまりのことであるから、そこに含まれる文書数分だけ重みを付けてベクトルの重心を求めている。

【００３１】次に、分類対象リスト３３からi とk とを取り除き、クラスタC を追加して、再度、分類対象リストの中で最も類似度の大きい組合せを抽出し、クラスタを生成する。

【００３２】こうして、類似度の最大値が所定の閾値を下回るまで、上記の過程を繰り返す。繰り返しが完了した時点で、他のクラスタの要素となっていないクラスタを列挙し、それぞれのクラスタに含まれる文書の集合を、類似文書集合とする。

【００３３】分類カテゴリ特徴量計算手段６は、図３に示すフローによって実現する。

【００３４】図３において、符号２，４，６は図１に対応し、６１は問合せ文書・回答文書対応管理部、６２は分類カテゴリ特徴ベクトル計算部、６３は分類カテゴリ特徴ベクトル、６４は形態素解析部を表わしている。

【００３５】問合せ文書・回答文書対応管理部６１は、各々の分類カテゴリに含まれる全ての回答文書について、該回答文書に対応する問合せ文書を回答文書分類カテゴリ４から取得する。該問合せ文書の各文を問合せ文書集合２および形態素解析部６４をへて単語に分解した状態で管理する。分類カテゴリ特徴ベクトル計算部６２は、以下の特徴ベクトルFVを求めて、これを以って分類カテゴリ特徴ベクトル６３とする。

【００３６】

6

【数１】

$$FV(i) = (w(i,1),\ldots,w(i,j),\ldots,w(i,N))$$

$$w(i,j) = \sum_{c=1}^{C(i)} tf(c,j) * \log (M / df(j))$$

【００３７】ここで、i は分類カテゴリの番号、N は問合せ文書集合における全単語数、C(i)は、分類カテゴリi に含まれる回答文書数、tf(c,j) は、分類カテゴリに含まれる回答文書c に対応する問合せ文書における単語j の出現回数、M は問合せ文書の総数、df(j) は、問合せ文書集合における単語j の出現回数である。

【００３８】分類カテゴリは、複数の類似する文書の集まりとして求められている。この分類カテゴリの特徴ベクトルを表すために、その中に含まれる文書の含む単語の情報を用いる。基本的には個別の文書の特徴ベクトルを求めるのと同じ考え方であるが、複数の文書を含んでいるから、各文書に含まれる単語の出現回数を全て足し合わせることによってカテゴリ全体の特徴ベクトルを求めている。

【００３９】適合カテゴリ選択手段８は、図４に示すフローによって実現する。

【００４０】図４において、符号４，７，８は図１に対応し、８１は形態素解析部、８２は新規問合せ文書特徴量計算部、８３は新規問合せ文書特徴ベクトル、８４は適合度計算部、８５は適合カテゴリリスト、８６はソート処理部を表わしている。

【００４１】形態素解析部８１は、新規に与えられた問合せ文書の各文を新規問合せ文書７から取り出して単語に分解する。

【００４２】新規問合せ文書特徴量計算部８２では、前記形態素解析部８１により抽出された単語情報に基づいて新規問合せ文書特徴ベクトル８３として以下の問合せ文書特徴ベクトルQVを算出する。

【００４３】QV＝(tf(1),...,tf(j),...,tf(N))

ここで、tf(j) は、新規問合せ文書における単語j の出現回数、N は問合せ文書集合における全単語数である。

【００４４】適合度計算部８４は、前記新規問合せ文書特徴量計算部８２により得られた特徴ベクトルと、前記分類カテゴリ特徴量計算手段６より得られた全ての分類カテゴリを回答文書分類カテゴリ４から抽出して当該分類カテゴリの特徴ベクトルとの適合度score を以下の計算式により計算して、適合カテゴリリスト８５を得る。

【００４５】

score(i)＝QV・FV(i) ／（｜QV｜・｜FV(i) ｜）

ここで、QVは、新規問合せ文書の特徴ベクトル、FV(i)は分類カテゴリi の特徴ベクトルである。

【００４６】新規問合せ文書と多くの単語が重なっている分類カテゴリを求めようとしている。そのため、QVを、一旦、単語からなるベクトルに分解した後にscore

7

を計算している。scoreは、各分類カテゴリの特徴ベクトルと新規問合せ文書の特徴ベクトルとの成す余弦として計算している。

【００４７】ソート処理部８６は、前記適合度計算部８４によって計算された適合度が高いものから順番に並べ替え、上位から所定の数の分類カテゴリを抽出し、検索結果として出力する。

【００４８】上記において、問合せ文書の分類装置について説明したが、当該問合せ文書の分類装置に対応する問合せ文書の分類方法を考慮することができ、かつまた当該問合せ文書の分類方法を記述したプログラムを用意しておいて当該プログラムを記録媒体に記録することができる。

【００４９】図５は本発明による問合せ文書の分類方法を表わす処理フローを示している。図５におけるステップ（Ｓ１）ないしステップ（Ｓ４）およびステップ（Ｓ６）ないしステップ（Ｓ９）は、上述の図１に示す符号１ないし符号４および符号６ないし符号９に示す各手段などの処理に対応している。即ち、

ステップ（Ｓ１）：回答文書集合１から個々の文書を抽出する。

【００５０】ステップ（Ｓ２）：問合せ文書集合２から文書を抽出する。

【００５１】ステップ（Ｓ３）：個々の文書について単語に分解し形態素解析を行って特徴ベクトルを計算した上で分類する。

【００５２】ステップ（Ｓ４）：分類した回答文書を保持する。

【００５３】ステップ（Ｓ６）：問合せ文書について特徴ベクトルを計算し、回答文書と対応づけて分類カテゴリ特徴量を計算する。そしてステップ（Ｓ４）に戻り保持する。

【００５４】ステップ（Ｓ７）：新規の問合せ文書を抽出する。

【００５５】ステップ（Ｓ８）：新規の問合せ文書について特徴ベクトルを計算し、ステップ（Ｓ６）において保持している分類カテゴリとの適合度を計算する。

【００５６】ステップ（Ｓ９）：適合したカテゴリを上位から所定の数だけ収集する。

【００５７】上記図５に示した問合せ分類方法はそれをプログラムの形で記述することができ、当該プログラムを記録媒体に記録することができる。したがって、本発明は、当該記録した記録媒体をも技術的範囲に含めることにする。

8

【００５８】

【発明の効果】以上説明した如く、本発明によれば、蓄積された問合せ・回答文書情報を、その内容に基づいて複数のカテゴリに分類し、新たに寄せられた問合せに対して、その内容が最も近いカテゴリを高い精度で選び出すことが可能となる。

【００５９】例えば、企業の問合せ応対窓口において、以前回答した事例と類似した問合せが寄せられた時に、前回に回答した時の事例を高精度に抽出できる。ただし、企業の問合せ対応窓口ではしばしば非常に大量の問合せに回答するし、類似する事例もごく大量存在する場合が有り得る。その時、類似する事例を片端から提示していったのでは作業者の負担になってしまうので、一旦類似する事例をカテゴリとしてまとめ、このカテゴリを検索結果として作業者に提示するようにしている。このことにより、作業者はより少ない数の候補から正しい情報を選び出すことができるので作業負担が軽減する効果が得られる。

【図面の簡単な説明】

【図１】本発明の問合せ文書分類装置のシステムブロック図である。

【図２】本発明の問合せ文書分類装置の回答文書分類手段の処理フローである。

【図３】本発明の問合せ文書分類装置の分類カテゴリ特徴量計算手段の処理フローである。

【図４】本発明の問合せ文書分類装置の適合カテゴリ選択手段の処理フローである。

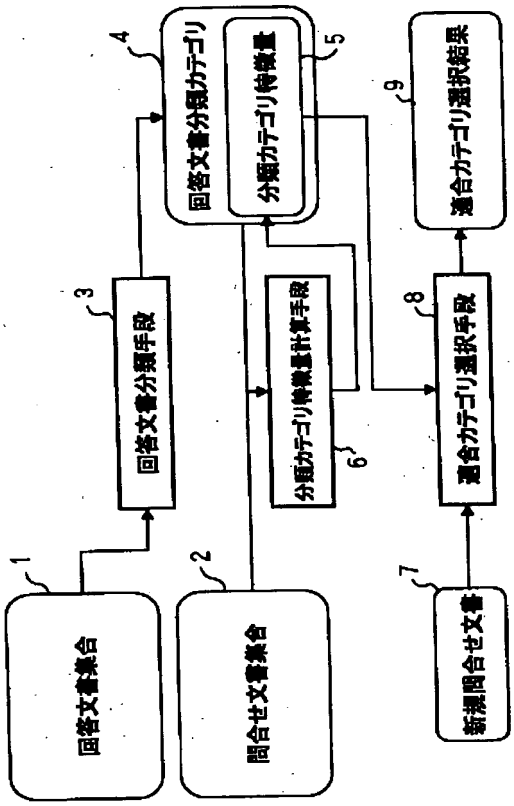
【図５】本発明の問合せ文書分類方法の処理フローである。

【符号の説明】

1　回答文書集合
2　問合せ文書集合
3　回答文書分類手段
4　回答文書分類カテゴリ
5　分類カテゴリ特徴量
6　分類カテゴリ特徴量計算手段
7　新規問合せ文書
8　適合カテゴリ選択手段
9　適合カテゴリ選択結果

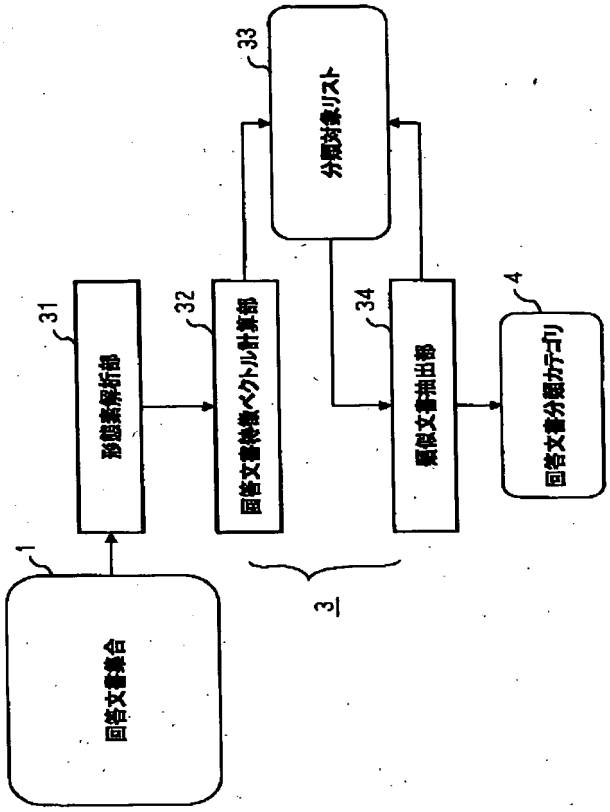
【図１】

実施例システムブロック図



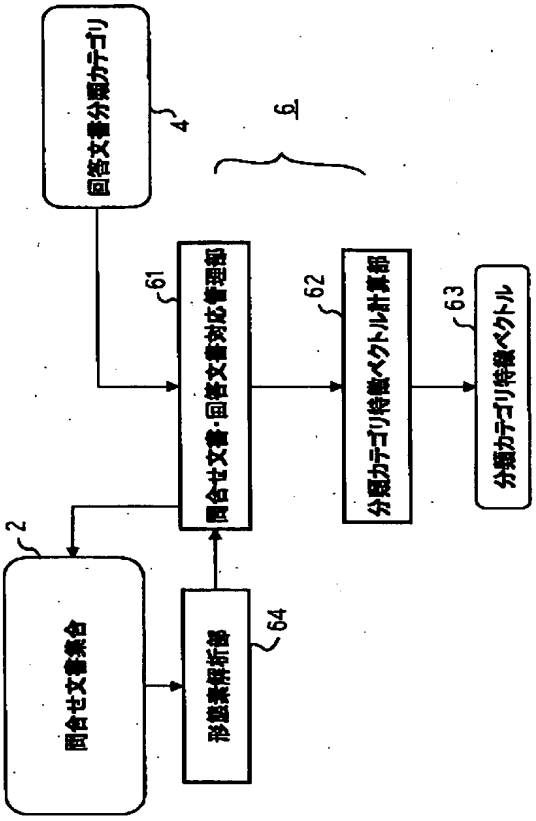

【図２】

回答文書分類手段の処理フロー



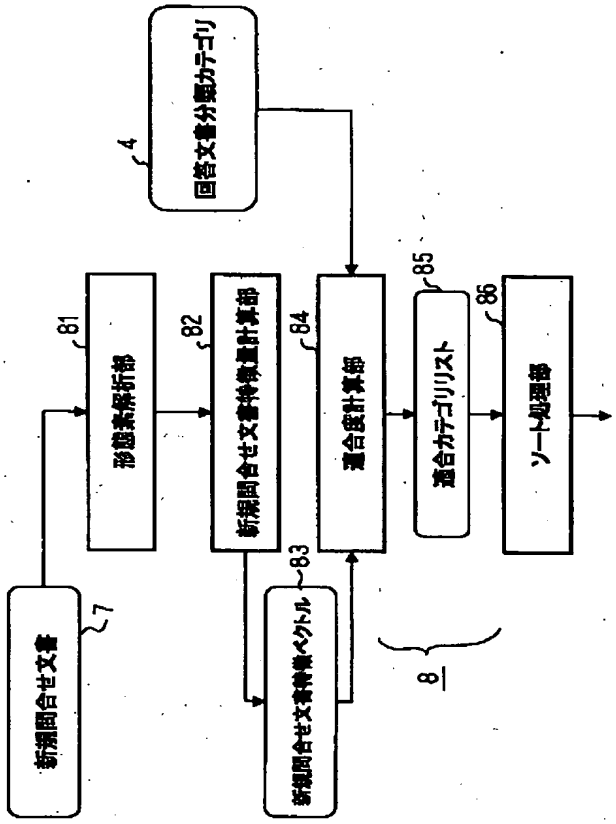
【図３】

分類カテゴリ特徴量計算手段の処理フロー

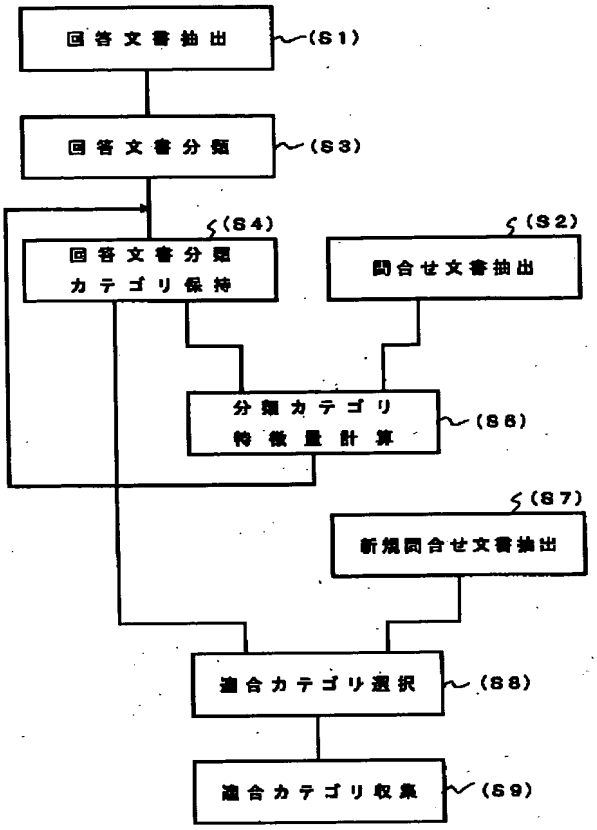



【図４】

適合カテゴリ選択手段の処理フロー



【図5】




フロントページの続き

(72)発明者　杉崎　正之
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　田中　一男
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

Fターム(参考)　5B075 ND03 ND36 NK06 NK32 UU06